M.P.H. THESIS

Linkage Disequilibrium Information within Monozygotic Twin Pairs

일란성 쌍둥이의 연관불균형 정보의 이용

FEBRUARY 2014

DEPARTMENT OF EPIDEMIOLOGY
GRADUATE SCHOOL OF PUBLIC HEALTH
SEOUL NATIONAL UNIVERSITY

Kim Jinseob

M.P.H. THESIS

Linkage Disequilibrium Information within Monozygotic Twin Pairs

일란성 쌍둥이의 연관불균형 정보의 이용

FEBRUARY 2014

DEPARTMENT OF EPIDEMIOLOGY
GRADUATE SCHOOL OF PUBLIC HEALTH
SEOUL NATIONAL UNIVERSITY

Kim Jinseob

Linkage Disequilibrium Information within Monozygotic Twin Pairs

일란성 쌍둥이의 연관불균형 정보의 이용

지도교수 성주헌

이 논문을 보건학석사 학위논문으로 제출함

2013 년 10 월

서울대학교 보건대학원

보건학과

김진섭

김진섭의 보건학석사 학위논문을 인준함

2013 년 12 월

위 원 장	김 호	(인)
부위원장	조성일	(인)
위 원	성주헌	(인)

Abstract

In conventional GWA studies, only one of monozygotic twin(MZ) pairs are considered because of same genetic information that we fomulate new method for gene-mapping study using resemblance information of phenotypes between MZ cotwins. Our new method used MZ concordance information as dependent variable(binomial trait) that Cochran-Amitage trend test(no covariate) or logistic regression(with covariate) was applied. Considering serum triglyceride(TG) or Hypertriglycedemia(hyperTG) traits and rs651821, well-known TG related SNP, power of our association test was calculated via creating simulated phenotypes and we apply our approach to 399 real MZ pairs in Healthy Twin Study, Korea and compare p-value with that of result of GWAS using 1,819 individuals including family structures. In result, rs651821 has some trend with hyperTG traits but is not significant in various hyperTG cutoffs as 120, 150, 180, 210, 240(range of p-value: 0.003-0.176). But compared with conventional GWAS in whole autosomal chromosome levels(537,158 SNPs), The authors find positive association in our new approach and GWAS's results using ROC curves and AUC values. Our methods using new information that concordance of MZ or probability of at least 1 MZ cotwin are affected can contribute gene mapping study.

Keywords: linkage disequilibium, monozygotic twin, genome-wide association study

Student Number: 2010-23807

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Recently genome-wide association studies have become a standard gene mapping method, but some limitations like Lack of well defined case and control groups, insufficient sample size, control for multiple testing problem, control for population stratification and "missing heritability" are issued[1, 2, 3].

Meanwhile, MZs have been either considered to be redundant or treated as unrelated individuals after randomly selecting one cotwin. It has been a common sense in academia that monozygotic twin (MZ) pairs, although they have unique strengths in detecting non-genomic etiology, do not contribute to gene mapping studies.

It is indeed true for linkage analysis, but resemblance of phenotypes between MZ cotwins does include information about linkage disequilibrium (LD) between the genetic markers and postulated disease-susceptibility loci (DSL) when the concordance/discordance rates or probability of twin has disease are compared across genotypes. For example, if a disease and disease susceptible loci(DSL) are highly penetrated in population, probability of both MZ are affected is in-

creased when number of disease suscetible alleles are increased (assuming additive model). But with low penetrance, above explanation isn't appropriated that we can probability of at least one of MZtwin is affected instead of probability of both MZ are affected.

This approach provides new information about DSL and contributes gene mapping stuidies, so Some GWAS limitations like power, control of multiple testing is improved via meta analysis and screening test[3, 4].

We attempted to formulate a method to detect association and suggest several ways of applying the information using hypertriglyceridemia as a model phenotype. At first, We find DSL using trivial GWAS study from the Healthy Twin Study, korea which do not utilize MZ information. And we perform power calculation of our method via creating MZ phenotype(eg, concordance or at least one is affected) of various scenario using DSL information and genotype data of founder in Healthy Twin Study. Last, We apply our methods to DSL and 399MZ in Healthy Twin Study that we compare results with trivial GWAS's result[5].

Chapter 2

Materials and Methods

2.1 Data

Healthy Twin Study, Korea is used in our study. This data includes 3,461 Korean families with 493 MZs and 537,158 SNPs in all autosomal chromosome genotyping with Affymetrix 6.0. We assume that genotypes of MZs are same and only one twin is genotyped.

2.2 Notation

It is natural to assume that if allele D (wild type allele is +) is in LD with true DSL, so genotypes are DD, D+ and ++. "A*A" is the proportion of MZ with both affected, "A*U" is the proportion of MZ with only one affected and "U*U" is the proportion of MZ with both unaffected. "Obs()" is observed rate and "prevalence" is prevalence rate among general population or study participants.

2.3 Our methods

We considered penetrance of diease and only considered additive effect model.

2.3.1 Method 1: Using Twin concordance rate - High penetrance model

Assuming disease is highly penetrated in population, if observed loci isn't associated with disease, probability of both MZ are affected is square of prevalence and probability of both MZ are not affected is square of 1-prevalence. In contrast, if observed loci has LD with DSL or is associated with disease, probability of both MZ are affected is increased and probability of both MZ are not affected are decreased with increasing number of disease susceptible allele. Manipulating this expression, we therefore obtain hypothesis that

$$
\begin{aligned}
&H_0 : Obs(A*A|DD) = Obs(A*A|D+) = Obs(A*A|++) = (prevalence)^2 \\
&Obs(U*U|DD) = Obs(U*U|D+) = Obs(U*U|++) = (1-prevalence)^2 \\
&H_1 : Obs(A*A|DD) > Obs(A*A|D+) > Obs(A*A|++) \\
&Obs(U*U|DD) < Obs(U*U|D+) < Obs(U*U|++)
\end{aligned}
\tag{2.1}
$$

We use only concordant MZ pairs, So hypothesis can be tested with Cochran–Armitage test for trend(no covariates) or logistic regression(with covariate) and formula's are

$$
T = \sum_{i=1}^{3} t_i (N_{1i} R_2 - N_{2i} R_1) \quad \text{(Test statistics for trend test)} \tag{2.2}
$$

	$++$	$D+$	DD	Sum
$A*A$	N_{11}	N_{12}	N_{13}	R_1
$U*U$	N_{21}	N_{22}	N_{23}	R_2
Sum	C_1	C_2	C_3	N

Table 2.1 Table for trend test: Method 1

t_i : weight(number of risk allele) for trend test ; $t_1 = 0, t_2 = 1, t_3 = 2$
T has chi-square distribution with 1 df(degree of freedom) assuming H_0.

$$logit(p(A*A)) = \alpha_0 + \alpha \cdot \text{covariate} + \beta \cdot \text{num(allele D)} + \epsilon \tag{2.3}$$

α_0: intercept, α: coefficient vector of covariates, β: coefficient of number of risk allele, num(allele D): number of risk allele, ϵ: error
β value is zero when H_0.

2.3.2 Method 2: Using probability of the affected exists within each MZ - Low penetrance model

If disease has low penetrance, similar to above, probability of the affected exist within each MZ isn't changed with increasing number of risk allele in null hypothesis and the probability is increased with increasing number of risk allele in alternative hypothesis. So expression is

$$H_0 : Obs(A*AorA*U|DD) = Obs(A*AorA*U|D+) = Obs(A*AorA*U|++)$$
$$H_1 : Obs(A*AorA*U|DD) > Obs(A*AorA*U|D+) > Obs(A*AorA*U|++) \tag{2.4}$$

	++	$D+$	DD	Sum
$A*AorA*U$	M_{11}	M_{12}	M_{13}	S_1
$U*U$	M_{21}	M_{22}	N_{23}	S_2
Sum	D_1	D_2	D_3	M

Table 2.2 Table for trend test: Method 2

We use all MZ pairs in method 2, hypothesis test is similar to above that

$$T = \sum_{i=1}^{3} v_i(M_{1i}S_2 - M_{2i}S_1) \quad \text{(Test statistics for trend test)} \tag{2.5}$$

v_i : weight(number of risk allele) for trend test ; $v_1 = 0, v_2 = 1, v_3 = 2$

$$logit(p(A*A \quad or \quad A*U)) = \alpha_0 + \alpha \cdot \text{covariate} + \beta \cdot \text{num(allele D)} + \epsilon \tag{2.6}$$

Formula notation and hypothesis test is same to above.

2.4 Conventinal GWAS to find DSL

We find rs651821 (p-value=4.5×10^{-15}) at chromosme 11 as DSL with triglyceride in whole family samples using FASTA method via R package "GenABEL" [6, 7]. This TG level-associated SNP is previously reported and we use rs651821 as DSL in our study[8, 9].

2.5 Simulation to calculate power

For performing simulations, we select 481 subset of people in Healthy Twin Study,Korea as a simulation dataset that are founder and non-missing at rs651821, assuming one people's genotype as of two persons of MZ. Next, we create various binary phenotypes for simulation based on simulation genotype data

and assuming causal loci and it's effect size as rs651821 and this SNP's effect size in "GenABEL". We consider various scenario with various heritability, prevalence, case:control ratio (heritability: $0.1, 0.2, \cdots, 0.8$; prevalence: $0.05, 0.1, 0.2, 0.3, 0.4$; proportion of case: $0.05, 0.1, 0.2, 0.3, 0.4, 0.5$) that we create 200 binary phenotypes for each scenario and calculate proportion of rs651821's p-value's under threshold(e.g 5×10^{-8}). Synthetic phenotypes are created by "GCTA" ver 1.20 and p-value's are calculated by "PLINK" ver 1.07[10, 11].

2.6 Apply to real MZ data

We apply above methods with 399 MZ pairs in Healthy Twin Study, Korea and consider various hypertriglycedemia disease cut-offs$(120, 150, 180, 210, 240)$ with "GenABEL" package with R 3.0.2, compare with previous GWAS's result. Finally, We compare our methods with GWAS in whole autosomal SNP level(537,158 SNPs) using ROC curves. We assume GWAS's p-values less than each significant levels $(10^{-3}, 10^{-4}, 10^{-5}, 10^{-6})$ as event and analyze ROC curve and area under curve(AUC) values with various cutoffs and methods.

Chapter 3

Results

3.1 Descriptive statistics of study populations

Descriptive informations of 399MZ pairs(157 male) in our study is presented in Table 3.1 that mean age is 38.87(s.d 7.31) and mean TG is 108.38(s.d 81.67). Number that Both MZ are affected is 71-7 according hyperTG cutoffs(120,150,180,210,240) and that of at least 1 MZ is affected is 163-44 according same cutoffs.

Next, We earn descriptive penetrance information via assuming penetrance as proportion of affected person(or MZ pairs) in population(or MZ pair) that have minor allele in rs651821 (Table 3.2). Considering hyperTG phenotype of individuals in MZ pairs(N=790), penetrance's are 28.9% to 5.6% according hyperTG cutoffs and those are 21.7%-1.9% and 40.1% 10.1% in high penetrance model and low penetrance model.

Variables		Mean ± S.D or N(%)
Age		38.87 ± 7.31
Sex	Male	157(39.3)
	Female	242(60.7)
TG		108.38 ± 81.67
Both MZ are affected (cutoff)	120	71(23.1)
	150	35(12.9)
	180	22(8.5)
	210	13(5.2)
	240	7(2.9)
At least 1 of MZ are affected (cutoff)	120	163(40.9)
	150	114(28.6)
	180	78(19.5)
	210	58(14.5)
	240	44(11.0)

Table 3.1 Descriptive statistics of 399 MZ pairs

Phenotype	Cutoff	Penetrance(%)
HyperTG(N=790)	120	28.9
	150	18.3
	180	11.5
	210	8.0
	240	5.6
High penetrance model	120 (N=303)	21.7
	150 (N=267)	12.0
	180 (N=254)	7.4
	210 (N=245)	3.9
	240 (N=239)	1.9
Low penetrance model (N=395)	120	40.1
	150	28.1
	180	18.3
	210	13.5
	240	10.1

Note: Penetrance is calculated as mean of that of 1 risk allele's and 2 risk alleles case

Table 3.2 Penetrance of various TG related phenotypes

3.2 Power calculation

Above to method, We calculate powers in various scenarios with 481 person's genotypes of rs651821 and synthetic phenotypes (Figure 3.1). We found obvious trend with heritability that mean power is 0.43, 0.76, 0.86, 0.92, 0.98, 0.997, 0.9998, 1 according 0.1-0.8 heritablity values that power is higher than 0.9 when heritability is greater than 0.4. Power is decreased (0.93-0.81) according prevalnece increased(0.05-0.4) and increased (0.61-0.96) with increasing case proportion (0.05-0.5).

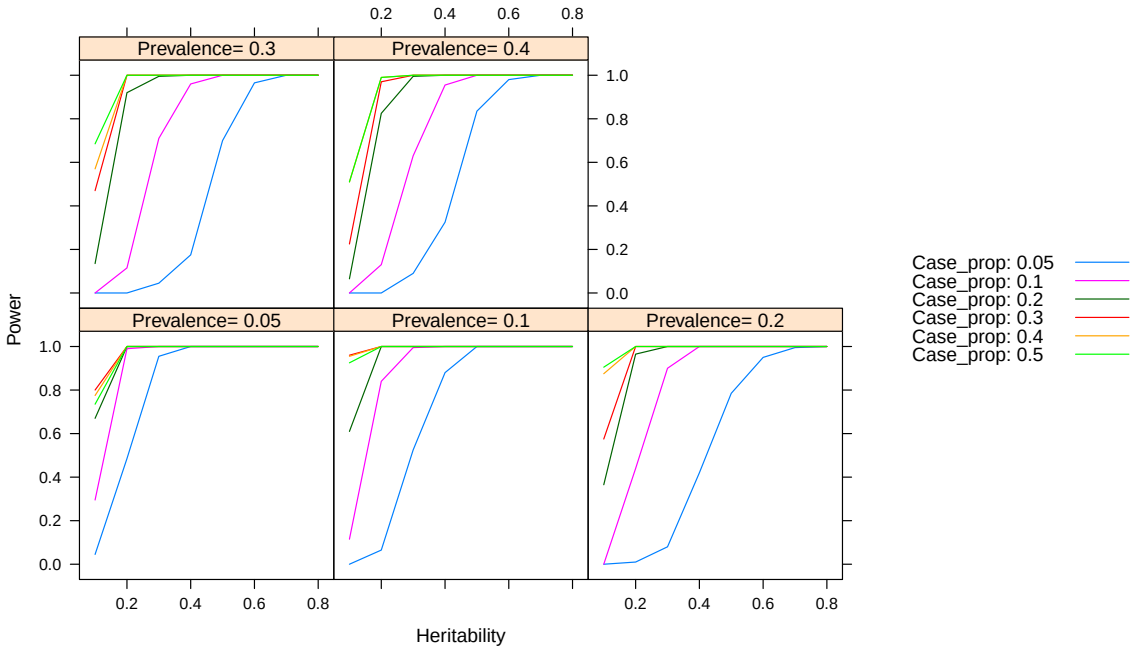

Figure 3.1 Results of power calculation using synthetic phenotypes

Model (SNP= rs651821)	Cutoff	P-value
GWAS(N=1819,trait=TG)		4.5×10^{-15}
High penetrance model	120	0.023
	150	0.011
	180	0.002
	210	0.024
	240	0.11
Low penetrance model	120	0.094
	150	0.176
	180	0.009
	210	0.003
	240	0.03

Note: "GWAS" is result of conventional GWAS

Table 3.3 Results of our Methods & meta-analysis

3.3 Apply to MZ data

We apply our method to 399 MZ pairs with various hyperTG cutoffs and that we find some trend between rs651821 information and various hyperTG traits(p-value's range: 0.002-0.176), but any significant results aren't found(significant level: 5×10^{-8}, Table 3.3).

3.4 Compare with conventional GWAS via whole chromosome

We compare p-values of 537,158 SNPs in GWAS and our methods using ROC curves (Figure 3.2). For various GWAS's significant levels $(10^{-3}, 10^{-4}, 10^{-5}, 10^{-6})$, We analyze ROC curve and area under curve(AUC) values with various cutoffs

and methods that low penetrance models have higher AUC values than high penetrance model(0.69 VS 0.61). There are no obvious trend according cut-off values (0.65,0.65,0.68,0.61,0.66 when cutoffs are 120,150,180,210,240) and AUC values are decreased(0.71,0.68,0.61,0.60) with increasing significant levels $(10^{-3}, 10^{-4}, 10^{-5}, 10^{-6})$.

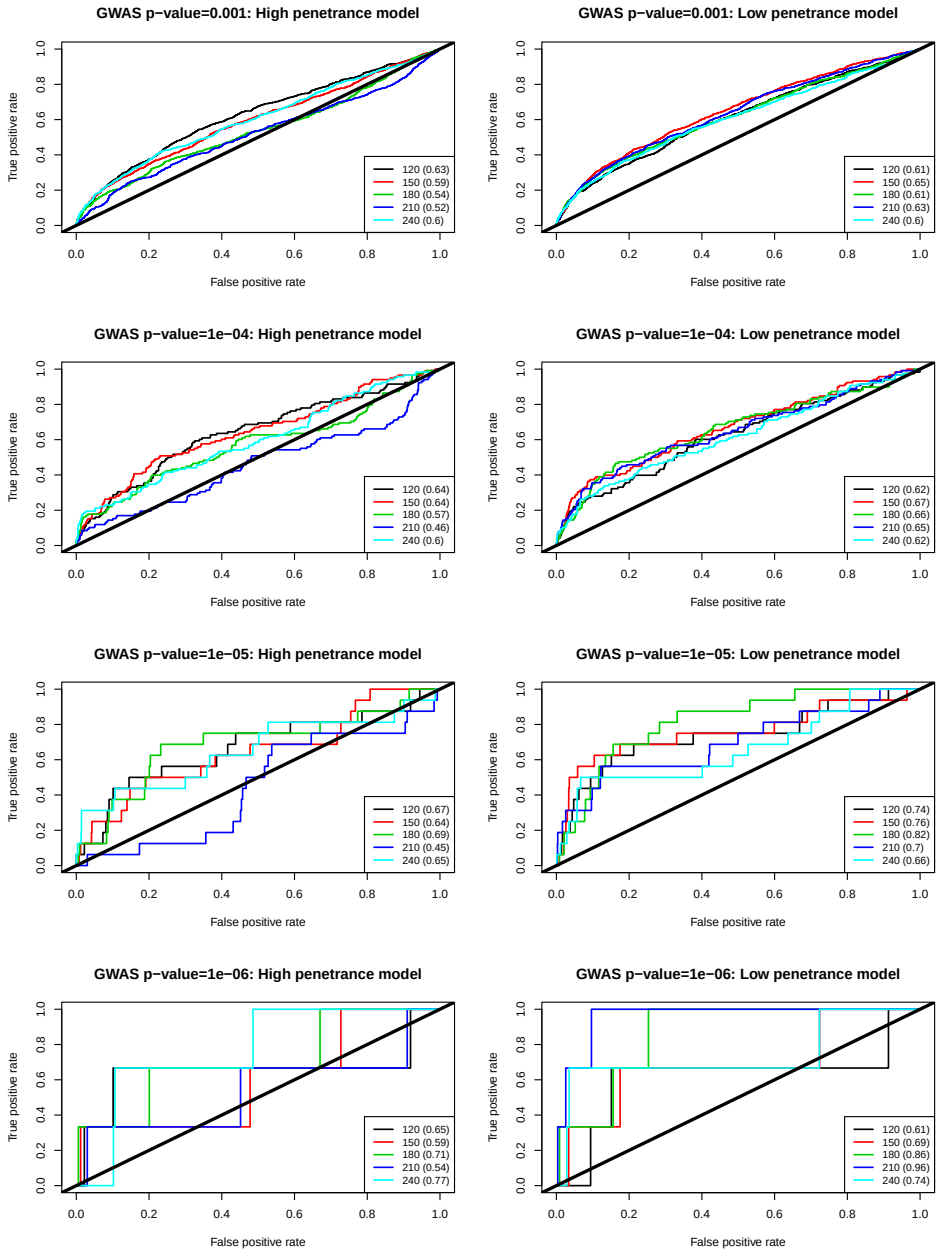

Figure 3.2 ROC curves and AUCs with various models and GWAS cutoffs

Chapter 4

Discussion

In this study, We analyze association of TG like traits and rs651821 with our method via power calculation, comparing with conventional GWAS's result, and compare our methods with GWAS in whole chromosome level.

We calculate TG's heritability as 0.4 using 1,819 individuals including family structure, this result and rs651821's significant p-value(4.5×10^{-15}) are compatible with power calculation that power is greater than 0.9 when heritability is greater than 0.4(Figure 3.1). But No sigificant results when applying our methods with various hyperTG cutoffs(Table 3.3), we think this is because of some limitations of our study. First, Defining heritability of phenotypes combining MZ information is very hard because our phenotypes are defined only in MZ. We think unmeasured heritability like status in our scenario is decreased because of combining phenotype informations of 2 MZ and this may affect to non-significant results. Second, information loss because of binomial transformation of trait that method of analyzing continuous trait is one of our next chalange. Making new phenotype similar to "Family Based Association

Test(FBAT)" is reasonable approach, but normality of new phenotype issue is problem, so further study is needed to solve this issue[12]. Next, our method's phenotype is differ to original binomial trait that new phenotypes in our high & low penetrance model(concordance information or probability at least one MZ affected) aren't direct measure of individual's TG levels or hyperTG status. In addition, case proportion and N are decreased when analyzing concordant MZ pairs(Table 3.1), or this is impossible to separate case of both MZ are affected with only one of MZ is affected in low penetrance model. Insufficient sample size and decreased sample due to combining 2 MZ's information can also affect non-significant results in various scenario. But most results of ROC curves using 537,158 SNPs in Figure 3.2 reveal that our approach is positively correlated with conventional GWAS, many of above issues will be overcome with sufficient sample size.

In spite of above limitaions, our findings show a way to extract LD information from MZ concordance or resemblance when one of cotwin was genotyped. Because the MZ resemblance is independent of other information conventionally used for GWAS, it can be combined with other results. The test can be also used to screen markers to alleviate the burden of multiple testing with appropriate threshold. Given that the MZ will be ever popularly applied for various omics studies, the addition of genomic information will facilitate multi-omics study in twin research.

Bibliography

[1] Lucia A Hindorff, Praveen Sethupathy, Heather A Junkins, Erin M Ramos, Jayashri P Mehta, Francis S Collins, and Teri A Manolio. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences*, 106(23):9362–9367, 2009.

[2] Thomas A Pearson and Teri A Manolio. How to interpret a genome-wide association study. *JAMA: the journal of the American Medical Association*, 299(11):1335–1344, 2008.

[3] Randall Johnson, George Nelson, Jennifer Troyer, James Lautenberger, Bailey Kessing, Cheryl Winkler, and Stephen O'Brien. Accounting for multiple comparisons in a genome-wide association study (gwas). *BMC genomics*, 11(1):724, 2010.

[4] John R Thompson, John Attia, and Cosetta Minelli. The meta-analysis of genome-wide association studies. *Briefings in bioinformatics*, 12(3):259–269, 2011.

[5] Joohon Sung, Sung-Il Cho, Yun-Mi Song, Kayoung Lee, Eun-Young Choi, Mina Ha, Jihae Kim, Ho Kim, Yeonju Kim, Eun-Kyung Shin, et al. Do

we need more twin studies? the healthy twin study, korea. *International journal of epidemiology*, 35(2):488–490, 2006.

[6] Yurii S Aulchenko, Stephan Ripke, Aaron Isaacs, and Cornelia M Van Duijn. Genabel: an r library for genome-wide association analysis. *Bioinformatics*, 23(10):1294–1296, 2007.

[7] Wei-Min Chen and Gonçalo R Abecasis. Family-based association tests for genomewide association scans. *The American Journal of Human Genetics*, 81(5):913–926, 2007.

[8] Sekar Kathiresan, Cristen J Willer, Gina M Peloso, Serkalem Demissie, Kiran Musunuru, Eric E Schadt, Lee Kaplan, Derrick Bennett, Yun Li, Toshiko Tanaka, et al. Common variants at 30 loci contribute to polygenic dyslipidemia. *Nature genetics*, 41(1):56–65, 2008.

[9] Aihua Tan, Jielin Sun, Ning Xia, Xue Qin, Yanling Hu, Shijun Zhang, Sha Tao, Yong Gao, Xiaobo Yang, Haiying Zhang, et al. A genome-wide association and gene–environment interaction study for serum triglycerides levels in a healthy chinese male population. *Human molecular genetics*, 21 (7):1658–1664, 2012.

[10] Jian Yang, S Hong Lee, Michael E Goddard, and Peter M Visscher. Gcta: a tool for genome-wide complex trait analysis. *The American Journal of Human Genetics*, 88(1):76–82, 2011.

[11] Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel AR Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul IW De Bakker, Mark J Daly, et al. Plink: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, 81(3):559–575, 2007.

[12] Steve Horvath, Xin Xu, and Nan M Laird. The family based association test method: strategies for studying general genotype–phenotype associations. *European journal of human genetics: EJHG*, 9(4):301–306, 2001.

요약

전장유전체 연관성 연구(Genome-wide assocation study, GWAS)가 질병관련 유전자를 찾는 표준 연구가 되어가고 있으나 일란성 쌍둥이는 그들의 완전히 일치하는 유전적 정보 때문에 오직 한명만이 연구대상이 될 수 있다. 이에 저자는 일란성 쌍둥이의 표현형의 일치여부 또는 둘 중 하나라도 질병에 걸린 사건을 새로운 표현형으로 하는 유전자 매핑 연구(gene mapping study) 방법을 제안한다. 본 연구에서는 일란성 쌍둥이의 표현형 일치여부를 독립변수로 하여 Cochran-Amitage trend test(공변량 없을 때) 또는 로지스틱 회귀분석(공변량 있을 때)을 이용하였다. 혈중 중성지방 또는 고중성지방혈증을 분석할 표현형으로, 혈청 중성지방과 연관성이 있는 단일염기다형성(SNP)인 rs651821을 대상으로 시뮬레이션을 이용하여 검정력(power)를 계산하였으며, 이를 한국 가족-쌍둥이 코호트의 399쌍의 일란성 쌍둥이와 그들의 다양한 고중성지방혈증 기준(120,150,180,210,240)에 따른 표현형에 적용하고 동일 코호트의 1,819명(가족구조 포함)을 이용한 일반적인 GWAS의 결과와 비교하였는데, rs651821과 고중성지방혈증이 관련성이 있는 것처럼 보이는 결과를 얻었으나 통계적으로 유의한 경우를 찾지 못하였다(p-values: 0.003-0.176). 그러나 가족-쌍둥이 코호트에서 분석한 537,158개의 SNP을 이용하여 일반적인 GWAS와 저자의 방법을 ROC curve와 area under curve(AUC) 값을 이용하여 비교해 보았을 때 양의 방향으로 비슷한 경향을 보임을 알 수 있었다. 본 연구에서 제시하고 있는 일란성 쌍둥이의 표현형의 정보를 이용한 분석 방법이 유전자 매핑 연구에 기여할 수 있을 것으로 기대한다.

주요어: 연관불균형, 일란성 쌍둥이, 전장유전체 연관성 연구
학번: 2010-23807

Acknowledgements

본 논문의 아이디어를 제공해주신 성주헌 교수님, 숙명여자대학교 통계학과의 김경아 교수님께 먼저 감사드립니다. 아이디어가 없었다면 본 논문이 나올 수 없었을 것입니다. 그리고 심사과정에서 많은 조언을 해 주신 김호 교수님, 조성일 교수님이 있었기에 본 논문의 내용이 더 알찬 내용으로 구성될 수 있었다고 생각합니다. 엉뚱한 말을 해도 받아주시고 냉철한 조언을 아끼지 않으신 김창엽교수님, 그리고 알찬 강의를 해 주신 보건대학원 교수님들, 많은 부분에서 같이 공부를 하면서 같이 실력을 늘려간 유전체역학교실원들, 예방의학교실 전공의 선생님들과 함께 했던 시간이 있었기에 제가 분석을 스스로 해내고 이렇게 글을 쓸 수 있었을 것입니다. 처음에는 무척 반대하셨으나 지금은 누구보다도 제 진로에 대해 이해하고 지지해주시는 부모님, 같이 연구자의 길을 가면서 힘이 되준 동생에게도 감사하다는 말씀을 드리고 싶으며, 무엇보다도 이 모든 과정을 잘 견뎌내고 마무리한 제 자신에게 수고했다고 칭찬과 격려를 하고 싶습니다. 서울대학교 보건대학원 석사과정에서 얻은 경험과 힘을 토대로 앞으로 더 발전하는 연구자가 되도록 하겠습니다. 감사합니다.